

# Variance on the Leaves of a Tree Markov Random Field: Detecting Character Dependencies in Phylogenies

Deeparnab Chakrabarty\*  
Microsoft Research, India  
Bangalore, 560 025  
Email: [dechakr@microsoft.com](mailto:dechakr@microsoft.com)

Sampath Kannan  
Dept. of Comp. and Inf. Science  
University of Pennsylvania  
Philadelphia, PA 19104  
Email: [kannan@cis.upenn.edu](mailto:kannan@cis.upenn.edu)

## Abstract

Stochastic models of evolution (Markov random fields on trivalent trees) generally assume that different characters (different runs of the stochastic process) are independent and identically distributed. In this paper we take the first steps towards addressing dependent characters. Specifically we show that, under certain technical assumptions regarding the evolution of individual characters, we can detect any significant, *history independent*, correlation between any pair of multistate characters. For the special case of the Cavender-Farris-Neyman (CFN) model on two states with symmetric transition matrices, our analysis needs milder assumptions. To perform the analysis, we need to prove a new concentration result for multistate random variables of a Markov random field on arbitrary trivalent trees – we show that the random variable counting the number of leaves in any particular subset of states has variance that is subquadratic in the number of leaves.

---

\*Work done as a postdoctoral researcher at the Dept of Comp. and Inf. Science, University of Pennsylvania

# 1 Introduction

Estimating the phylogeny or evolutionary history of a set of organisms is an important problem in biology [18, 10]. The problem has the following general form: data is available about the species alive today, and the goal is to find a tree with these species at the leaves that “best explains” the data. Specific formulations of the problem are arrived at by specifying the type of data that is used and the notion of the tree that best explains this data. While early attempts at phylogeny construction used morphological characters, the data these days is derived by and large from biomolecular sequences such as protein and DNA sequences. Homologous sequences with presumed common evolutionary origin are observed for each of the species of interest. These sequences are aligned (as well as possible) in a multiple sequence alignment. Each position in such an alignment is called a character and the values taken on by these characters are called its states. Character-based methods such as parsimony seek a tree, and states at each internal node for each character, to minimize the number of state changes among the characters over all the edges of the tree. Distance-based methods convert this raw data into pairwise distances between species and seek an edge-weighted tree such that the input distances best fit the distances in the tree under various norms.

Arguably, the most principled approach to phylogeny estimation is to select a family of stochastic models for character evolution, and find parameters for a model within this family that is most likely to have generated the data [9, 12]. Such stochastic models can be viewed as Markov random fields on rooted trivalent trees. The root has a state drawn from some distribution. Each node transmits its state to its two children via a Markovian process. For each character  $i$ , along a parent-child edge  $e = (u, v)$ , there is a transition matrix  $M_{e,i}(a, b)$  that defines for each pair of states  $a$  and  $b$  the probability that  $v$  is in state  $b$  given that  $u$  is in state  $a$ . When the character is clear from the context we refer to this matrix as  $M_e$  and when the edge is also clear we refer to it simply as  $M$ . The unknown parameters are the shape of the tree and the transition matrices on each edge. The goal of statistical methods is to infer these parameters in order to maximize the probability of generating the given data. Many specific families of models have been considered. Among the simplest are two-state, symmetric models, called the Cavender-Farris-Neyman (CFN) [17, 7, 1] models, where on any edge  $e$ , each character has the same, symmetric transition matrix  $M_e$ .

Under these models, considerable work [8, 4, 5, 16, 15, 3] has been done to determine the number of characters needed to infer the phylogenetic tree under reasonable technical constraints on the transition matrices. Almost all of these works assume that the stochastic processes of the various characters are independent and identically distributed. This assumption might not be biologically realistic. Changes at one position of a DNA sequence or amino acid sequence are likely to be correlated with changes at other positions because of constraints on size, charge, hydrophobicity, etc. of the molecules involved. Thus we need to begin to understand how to infer both the evolutionary tree and the dependence relationship between characters when we drop the i.i.d. assumption. Of course, allowing arbitrary dependence between characters can lead to problems where there is insufficient information to reconstruct the tree. We need to carefully choose a dependence model that is tractable, yet realistic. In the literature no general procedures with provable properties have been proposed for detecting dependence. There are only simple, heuristic statistical approaches. (See, for example, [14] and the references cited therein.) In this paper we take the first steps along these lines: we suggest a simple model of *preferred state* dependence. In this model, two dependent characters are biased towards evolving to certain pairs of states in comparison to how they would behave if they were independent. We assume this bias to be uniform, and irrespective of the initial states of these characters. Under certain technical conditions we show that it is possible to infer

the (pairwise) dependence structure among various characters.

*Informal Statement of Results.* We make the following technical assumption on the transition matrix  $M$  of the various Markov processes: the norm  $\|M\| \triangleq \max_{\mathbf{0} \neq x \perp \mathbf{1}} \|x^T M\|_1 / \|x\|_1$  is upper bounded by  $\lambda < 1$ . The assumption implies that the corresponding Markov chain is rapidly mixing.

As stated above, the general stochastic model of evolution assumes a different transition matrix for every edge and character. If two characters are independent, then the joint transition matrix governing their evolution is the tensor product of the individual transition matrices. If two characters are dependent, then our model assumes that the joint transition matrix on any edge is significantly “far away” from the tensor product. Furthermore, we assume that this correlation is *uniform* across all edges, and *history independent*, that is, it is same irrespective of the initial state of the two characters. Formally, the difference between the two joint transition matrices, dependent and independent, is rank one. See Section 2 for precise definitions.

The main tool we develop in this paper is a concentration bound for tree Markov random fields. Let  $Z$  be the random variable counting the number of occurrences of a character in a particular subset of states at the leaves of a rooted tree. We show that as long as  $\|M_e\| \leq \lambda < 1$  for each edge  $e$ ’s transition matrix, and the tree is trivalent (internal nodes have degree 3), the variance of  $Z$  is sub-quadratic in the number of leaves. For instance, this implies by Chebyshev’s inequality, that  $Z$  is concentrated around its mean if the latter is of the order of the number of leaves.

In the literature in various fields (computational biology [15], communications [6], statistical physics [11, 19]), concentration bounds have been derived for a similar random variable: for 2 state characters with states encoded as  $\{-1, +1\}$  (instead of  $\{0, 1\}$  as we do), and each edge having a symmetric transition matrix (that is, the CFN model), the random variable is the sum of the leaf states. It is known [15] that there is a threshold mutation probability that decides whether the ratio of variance to squared mean is bounded or not. Note that we do not observe such a phase transition; this is not surprising since our random variable is shifted additively. Nonetheless, our concentration result is exactly what we need for our application. Furthermore, our result holds for any number of states, and possibly will have other applications besides this work. We use the above result to get the following results on detecting dependence.

- For multistate characters, we show that if the joint transition matrix of two independent characters has bounded norm ( $< 1/2$ ), then there exists an algorithm to detect dependence.
- For the special case of the two state CFN model, we need the much milder assumption of  $\|M_{e,i}\| < 1$  for detecting dependence.

Arguably, the assumption made in our second result above is biologically feasible and indeed has been made explicitly or implicitly in existing literature [2, 8, 4, 5]. We believe the first assumption is a technicality, and conjecture that the analysis can be made to go through making only the weaker assumption of  $\|M_{e,i}\| < 1$ . This requires understanding a Markov chain that changes its transition matrix at each step, in particular, the relation between the stationary vector of this chain and the stationary vectors of the different transition matrices at each step.

## 2 Preliminaries

*Tree Markov Random Fields.* In a (rooted) tree Markov random field, we have a rooted tree  $T$  with root  $r$ . We assume in this paper that the tree is trivalent, that is, every internal node has degree

3 and the root has degree 2. This suffices for the application at hand, although our result holds more generally. Each node  $v \in V(T)$  has an associated binary random variable  $X_v$  taking values in a state space  $\mathcal{S}$  of size  $|\mathcal{S}| = s$ . Each edge  $e \in E(T)$  has an associated  $s \times s$  transition matrix  $M_e$ . In the ensuing stochastic process, the root random variable  $X_r$  is associated with an arbitrary distribution over  $\mathcal{S}$ . For every edge  $e = (u, v)$  with  $u$  as parent of  $v$ , we have

$$\mathbf{P}[X_v = b | X_u = a] = M_e(a, b)$$

for all  $a, b \in \mathcal{S}$ . Given a matrix  $M$ , define the following norm

$$\|M\| := \max_{\mathbf{0} \neq x \perp \mathbf{1}} \frac{\|x^T M\|_1}{\|x\|_1}$$

Observe that the norm is the maximum variational distance between the distributions induced by taking one step of the corresponding Markov chain, starting from two distinct distributions over the state space. The maximizing  $x$  can be thought of as  $p - q$  for two probability distributions  $p, q$  with *disjoint* support implying  $\|x\| = 2$ ; the observation follows since the numerator is twice the variational distance. Henceforth, we assume that all transition matrices in this paper have  $\|M\| \leq \lambda < 1$ . In other words, all the corresponding Markov chains are rapidly mixing.

Given a subset of states  $A \subseteq \mathcal{S}$ , let  $Z_A$ , or simply  $Z$ , denote the number of leaves of the tree  $T$  whose state belongs to  $A$ . In Section 3, we upper bound the variance of this random variable.

**Theorem 1.** *Given an  $n$  leaf, trivalent tree  $T$ , and a subset  $A \subseteq \mathcal{S}$ ,  $\mathbf{V}(Z_A) = O(n^{2-2\log_2(1/\lambda)})$ .*

*Phylogenetic Trees.* For maximum likelihood methods, the process of evolution is typically modeled as a tree Markov random field. The  $n$  leaves of the tree denote extant species and the interior vertices represent speciation events. The root is the common ancestor of the  $n$  species.

Each node of the tree is associated with a (aligned) sequence of  $k$  *characters* denoted as  $[k]$ , which evolve down the tree. Each character takes a value from a certain set of *states*  $\mathcal{S}$  such as  $\{0, 1\}$ ,  $\{A, C, G, T\}$ ,  $\{20 \text{ amino acids}\}$ , etc. The size of  $\mathcal{S}$  is assumed to be an arbitrary but finite constant. The random variable  $X_i(v) \in \mathcal{S}$  is used to denote the state of character  $i$  at this vertex  $v$ . In general, each character  $i$  and edge  $e = (u, v)$  of the tree has an associated  $s \times s$  transition matrix  $M_{e,i}$ , where  $\forall a, b \in \mathcal{S}, M_{e,i}(a, b) = \mathbf{P}[X_i(v) = b \mid X_i(u) = a]$ . Specific models assume that the edge transition matrices are drawn from specific families of matrices. A simple, popular model is the Cavender-Farris-Neyman (CFN) model where the characters are binary (with states 0 and 1) and the transition matrix is *symmetric*, with equal probabilities for state changes (mutations) from 0 to 1 and 1 to 0. That is,  $M_e[0, 1] = M_e[1, 0] = p_e$ . Thus, in this model, transition matrices are described by a single scalar parameter per edge.

As mentioned in the introduction, most of the literature assumes that every character evolves in an i.i.d fashion. In this work, we are interested in detecting dependencies between pairs of characters, if any. We next define our model of dependence.

*Dependence Model.* Fix a pair of characters  $i, j \in [k]$ . Let  $X$  be the ‘compound character’ with  $X(u)$  on vertex  $u$  being the ordered pair  $(X_i(u), X_j(u))$ . Thus,  $X$  is defined on a state set of size  $s^2$ . If  $X_i$  and  $X_j$  are independent<sup>1</sup>, then observe that  $X$  also evolves as a Markov random field, with the  $s^2 \times s^2$  transition matrix at edge  $e$  defined as  $\hat{M}_e := M_{e,i} \otimes M_{e,j}$ . More precisely, the  $s^2$  states are thought of a ordered pair of states, and  $(\hat{M}_e)_{(a,b),(a',b')} := M_{e,i}(a, a') \cdot M_{e,j}(b, b')$

---

<sup>1</sup>For brevity, henceforth, we will alternately say  $i$  and  $j$  are (in)dependent, to mean  $X_i$  and  $X_j$  are (in)dependent.

In our model of dependence, we assume that for two dependent characters, one can still define an  $s^2 \times s^2$  matrix  $\hat{M}_e$  for every edge  $e$  such that the evolution of the compound character can be described as a tree Markov random field with parameters  $\hat{M}_e$ . Furthermore, if  $i$  and  $j$  are dependent, we assume there is *significant, uniform* correlation which is *history independent*. To make this precise, consider any edge  $e = (u, v)$  in the tree and fix a pair of states  $(a, b) \in \mathcal{S} \times \mathcal{S}$  for  $(X_i(u), X_j(u))$ . Note that  $\hat{M}_e$  induces a probability distribution over  $\mathcal{S} \times \mathcal{S}$  at vertex  $v$ . Call this  $P_{\hat{M}}^{(a,b)}(v)$ . If  $i$  and  $j$  were independent, then call the probability distribution at  $v$ ,  $P_{M_i \otimes M_j}^{(a,b)}(v)$ . The ‘drift’ caused by dependence is precisely the vector  $\delta_e^{(a,b)} := P_{\hat{M}}^{(a,b)}(v) - P_{M_i \otimes M_j}^{(a,b)}(v)$ .

We assume that if  $i$  and  $j$  are dependent, then there exists a constant  $\delta > 0$ , such that:

$$\begin{aligned} \forall (a, b) \in \mathcal{S} \times \mathcal{S}, \quad \|\delta_e^{(a,b)}\|_1 &\geq 2\delta && \text{(Significant Correlation)} \\ \forall e, e' \in E, (a, b) \in \mathcal{S} \times \mathcal{S}, \quad \delta_e^{(a,b)} &= \delta_{e'}^{(a,b)} && \text{(Uniform Dependence)} \\ \forall (a, b), (a', b') \in \mathcal{S} \times \mathcal{S}, \quad \delta_e^{(a,b)} &= \delta_e^{(a',b')} && \text{(History Independence)} \end{aligned}$$

In other words, there is a vector  $\delta \in \mathbb{R}^{\mathcal{S} \times \mathcal{S}}$ ,  $\|\delta\|_1 \geq 2\delta$  and  $\sum_i \delta(i) = 0$ , such that, we have  $\hat{M}_e - M_{e,i} \otimes M_{e,j} = \mathbf{1}\delta^T$  for each edge  $e$ , and dependent characters  $i, j$ .

Given the dependence model above, an algorithm detects dependence if given any two characters  $i$  and  $j$  it can detect whether they are independent or dependent as per our model, with high probability (whp). By whp, we mean that the failure probability should be an inverse polynomial of the number of leaves. For general multistate characters, we give an algorithm for detecting dependence if the norm of each joint transition matrix is upper bounded away from  $1/2$ .

**Theorem 2.** *If  $\|M_{e,i} \otimes M_{e,j}\| < 1/2$  for all edges, then there exists an algorithm to detect dependence for multiple finite state characters.*

*Special case: CFN Model.* Recall in the CFN model that the transition matrix of each character on edge  $e$  is  $M_e = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}$  where  $p := p_e$ , is the probability of mutation on edge  $e$ . For dependent characters  $i$  and  $j$ , we assume there exists for each edge a 4-dimensional vector  $\delta$  such that  $\|\delta\|_1 \geq \delta$  (significant correlation) and  $\hat{M}_e = M_{e,i} \otimes M_{e,j} + \mathbf{1}\delta^T$  (history independence). Note that since both  $\hat{M}$  and  $M_i \otimes M_j$  are probability matrices, we have  $\sum_{i=1}^4 \delta(i) = 0$ . For this model, we can detect dependence with a much weaker assumption.

**Theorem 3.** *In the two state CFN model, there exists an algorithm to detect dependence between any two characters as long as  $\|M_{e,i}\| < 1$  for all edges.*

*Comments on Assumptions.* If there is no significant correlation between two dependent characters, then one cannot expect to detect dependence. Furthermore, ‘low dependence’ may not cause problems for algorithms that assume complete independence. The uniform dependence assumption can be relaxed to *consistent* dependence: if on a certain edge two dependent characters give a higher probability to the joint state  $(a, b)$  than they would have if they were independent, then they should prefer it, although maybe not to the same degree, on every edge. However, to keep the calculations clean we assume uniformity. Consistency seems to be a biologically resonable assumption. It is also infeasible to detect dependence with inconsistent characters with data just from the leaves. History independence is based on a memoryless principle for bias. This is a stronger (biological) assumption than the other two, however, it seems essential for this work, and we leave it as a challenge to relax or remove this.

### 3 Bounding the Variance

In this section, we prove Theorem 1. Fix any subset  $A \subseteq \mathcal{S}$ . Recall that  $Z$  is the random variable counting the number of leaves of  $T$  in a state in  $A$ . More generally, we define  $Z(T_u)$  to be the number of leaves in the sub-tree of  $T$  rooted at  $u$  in a state in  $A$ . Recall  $\lambda < 1$  is an upper bound on the norm of any of the transition matrices.

**Theorem 4.** (Restatement of Theorem 1) *Given any  $n$  leaf trivalent tree  $T$ ,  $\mathbf{V}(Z) = O(n^{2-2\log_2(1/\lambda)})$ .*

*Proof.* For any vertex  $v$  in  $V(T)$ , let  $L_v$  denote the leaves in the subtree rooted at  $v$ , and let  $Z_v$  denote  $Z(T_v)$ . Fix a leaf  $\ell$  and let  $e_1, e_2, \dots, e_t$  be the edges on the path from the root to  $\ell$ . Let  $M$  denote the matrix  $M_{e_t} \cdot M_{e_{t-1}} \cdots M_1$ ; this is the transition matrix from root to leaf  $\ell$ . In particular,  $M$  is row-stochastic (row entries add up to 1). We state a couple of facts about row stochastic matrices. The lemma below was suggested to us by Nikhil Srivastava [20], and a similar lemma can be found as Lemma 4.12 in [13].

**Lemma 1.** *For any two row stochastic matrices  $M_1$  and  $M_2$ , we have  $\|M_1 M_2\| \leq \|M_1\| \cdot \|M_2\|$ .*

*Proof.* Note that given any  $x \perp \mathbf{1}$ , and any row stochastic matrix  $M$ , the vector  $y = x^T M$  also satisfies  $y \perp \mathbf{1}$ . This is because  $M\mathbf{1} = \mathbf{1}$ . If  $x$  is the vector achieving the maximum  $\|x^T(M_1 M_2)\|_1 / \|x\|_1$ , then we get that

$$\|M_1 M_2\| = \frac{\|x^T(M_1 M_2)\|_1}{\|x\|_1} = \frac{\|x^T(M_1 M_2)\|_1}{\|x^T M_1\|_1} \cdot \frac{\|x^T M_1\|_1}{\|x\|_1} \leq \|M_2\| \cdot \|M_1\|$$

□

Lemma 1 implies that  $\|M\|$  is at most  $\lambda^t$ . In particular, this shows for a leaf  $\ell$  at a distance  $t$  from a root vertex  $r$ , and for any states  $b, c \in \mathcal{S}$ , we have

$$|\mathbf{P}[X_\ell \in A | X_r = b] - \mathbf{P}[X_\ell \in A | X_r = c]| \leq \|x^T M\|_1 \leq 2\lambda^t \quad (1)$$

where  $x$  above is the vector with  $x_b = 1$ ,  $x_c = -1$ , and  $x_s = 0$  otherwise.

Given a vertex  $v$  and two states  $b, c \in \mathcal{S}$ , let  $\Delta_v(b, c) := |\mathbf{E}(Z_v | X_v = b) - \mathbf{E}(Z_v | X_v = c)|$ . Expanding,

$$\begin{aligned} \Delta_v(b, c) &= \left| \sum_{\ell \in L_v} (\mathbf{P}[X_\ell \in A | X_v = b] - \mathbf{P}[X_\ell \in A | X_v = c]) \right| \\ &\leq \sum_{\ell \in L_v} |\mathbf{P}[X_\ell \in A | X_v = b] - \mathbf{P}[X_\ell \in A | X_v = c]| \end{aligned}$$

Using (1), we have  $|\mathbf{P}[X_\ell \in A | X_v = b] - \mathbf{P}[X_\ell \in A | X_v = c]| \leq 2\lambda^{d(v, \ell)}$ , implying

$$\Delta_v(b, c) \leq \Delta(v) \triangleq 2 \sum_{\ell \in L_v} \lambda^{d(v, \ell)}, \quad \forall v, b, c \quad (2)$$

We now bound the variance of  $Z$  as a function of the  $\Delta(v)$ 's.

**Lemma 2.**  $\mathbf{V}(Z) \leq \frac{1}{2} \sum_{v \in V(T) \setminus r} \Delta^2(v)$

*Proof.* Recall that  $T_u$  denotes the subtree of  $T$  rooted at  $u$  and  $Z_u = \sum_{\ell \in L(T_u)} \mathbf{1}_{X_\ell \in A}$ . We now show using induction on the height that  $\mathbf{V}(Z_u|X_u = b) \leq \frac{1}{2} \sum_{v \in V(T_u) \setminus u} \Delta^2(v)$  for any vertex  $u$ , and any  $b \in \mathcal{S}$ . Note that the claim is vacuously true when  $u$  is a leaf.

Let  $u$  have children  $v_1, \dots, v_k$  (if the tree is binary,  $k = 2$ , but this lemma holds for any tree). Assume we have proved the inductive claim for the  $v_i$ 's. Note that conditioned on  $X_u$ , the random variables  $Z_{v_1}, Z_{v_2}, \dots$  are independent, since they are on different subtrees. Therefore, for any  $b \in \mathcal{S}$ ,  $\mathbf{V}(Z_u|X_u = b) = \sum_{i=1}^k \mathbf{V}(Z_{v_i}|X_u = b)$ .

We now show that for any parent-child pair  $e = (u, v)$  and any state  $b \in \mathcal{S}$ , we have

$$\mathbf{V}(Z_v|X_u = b) = \sum_{s \in \mathcal{S}} P_{bs} \mathbf{V}(Z_v|X_v = s) + \frac{1}{2} \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} \Delta_v^2(s, s') \quad (3)$$

where  $P_{bs} = \mathbf{P}[X_v = s|X_u = b]$ .

We introduce some notational shorthand just to keep the calculation below simple. We forgo the subscript on  $Z_v$ , let  $V := \mathbf{V}(Z|X_u = b)$ , use “ $u = s$ ” to imply  $X_u = s$ , and use  $\mathbf{E}^2[Z]$  to denote  $(\mathbf{E}[Z])^2$ . Now, by definition,  $V = \mathbf{E}[Z^2|u = b] - \mathbf{E}^2[Z|u = b]$ . The first term evaluates to

$$\mathbf{E}[Z^2|u = b] = \sum_{s \in \mathcal{S}} P_{bs} \mathbf{E}[Z^2|v = s]$$

The second term evaluates to

$$\mathbf{E}^2[Z|u = b] = \left( \sum_{s \in \mathcal{S}} P_{bs} \mathbf{E}[Z|v = s] \right)^2 = \sum_{s \in \mathcal{S}} P_{bs}^2 \mathbf{E}^2[Z|v = s] + \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} \mathbf{E}[Z|v = s] \mathbf{E}[Z|v = s']$$

Observing  $P_{bs}^2 = P_{bs} - P_{bs}(1 - P_{bs})$ , we get  $V =$

$$\sum_{s \in \mathcal{S}} P_{bs} (\mathbf{E}[Z^2|v = s] - \mathbf{E}^2[Z|v = s]) + \sum_{s \in \mathcal{S}} P_{bs}(1 - P_{bs}) \mathbf{E}^2[Z|v = s] - \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} \mathbf{E}[Z|v = s] \mathbf{E}[Z|v = s']$$

The first term above is the first term in the RHS of (3). Furthermore, noting that  $P_{bs}(1 - P_{bs}) = \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'}$  since  $P_{bs}$ 's sum up to 1, we get that the second two terms is

$$\frac{1}{2} \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} (\mathbf{E}^2[Z|v = s] + \mathbf{E}^2[Z|v = s'] - 2\mathbf{E}[Z|v = s] \mathbf{E}[Z|v = s']) = \frac{1}{2} \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} \Delta_v^2(s, s')$$

which establishes (3). By induction, the first term in the RHS is at most  $\frac{1}{2} \sum_{x \in V(T_v) \setminus v} \Delta^2(x)$ . Since  $\Delta_v(s, s') \leq \Delta(v)$ , we get that the second term is at most  $\Delta^2(v) \sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} / 2 \leq \frac{1}{2} \Delta^2(v)$ . The last inequality follows from the fact  $\sum_{s \neq s' \in \mathcal{S}} P_{bs} P_{bs'} \leq (\sum_{s \in \mathcal{S}} P_{bs})^2 = 1$ .  $\square$

We now bound  $\sum_{v \in V(T) \setminus r} \Delta^2(v)$  for any  $n$  leaf trivalent tree. Recall,  $\Delta(v) = 2 \sum_{\ell \in L_v} \lambda^{d(v, \ell)}$ . The following claim bounds  $\Delta(u)$  for any  $u$  in a binary tree.

**Claim 1.** *For any vertex  $u$  with  $n$  leaves in its subtree,  $\Delta(u) = O(n^{1-\eta})$  where  $\eta = \log_2(1/\lambda)$ .*



*Proof.* Note that  $\Delta(u) = 2 \sum_{i \geq 1} |L_i| \lambda^i$  where  $L_i$  is the set of leaves at a distance  $i$  from  $u$ . Since  $\lambda < 1$ , an  $n$  leaf tree which maximizes  $\Delta(u)$  will make the tree as balanced in height as possible. (This can be proved by a “swapping” argument similar to the proof of optimality of Huffman trees.) In particular, the maximizing tree has all leaves at distance  $\lfloor \log n \rfloor$  or  $\lfloor \log n \rfloor + 1$ . Therefore,  $\Delta(v) \leq \frac{2}{\lambda} \cdot n \lambda^{\log n} = \frac{2}{\lambda} \cdot n^{1 - \log(1/\lambda)}$ .  $\square$

**Lemma 3.**  $\mathbf{V}(Z) = O(n^{2-2\log(1/\lambda)})$ .

*Proof.* Let  $\mathcal{V}(n)$  be a function that denotes the maximum value of  $\sum_{v \in V(T) \setminus r} \Delta^2(v)$  over all  $n$ -leaf binary trees. Now given an  $n$ -leaf tree  $T$ , let  $u$  be the *centroid* of  $T$ . That is,  $n/3 \leq |L_u| \leq 2n/3$ . It is easy to see this is well defined. Let  $T_u$  denote the subtree of  $T$  rooted at  $u$ , and let  $T'_u$  denote the subtree of  $T$  with all descendants of  $u$  deleted. Note that both  $T_u$  and  $T'_u$  are binary trees, and have  $\rho n$  and  $(1 - \rho)n$  leaves for  $\rho \in [1/3, 2/3]$ . By induction, we may assume

$$\sum_{v \in V(T_u) \setminus u} \Delta^2(v) \leq \mathcal{V}(\rho n), \quad \text{and} \quad \sum_{v \in V(T'_u) \setminus r} \Delta^2(v) \leq \mathcal{V}((1 - \rho)n)$$

Suppose  $u = u_0, u_1, \dots, u_r = r$  is the unique path from  $u$  to  $r$  in  $T$ . Note that the  $\Delta(v)$ 's in tree  $T'_u$  are the same as in tree  $T$  for all vertices except the  $u_i$ 's. For each  $u_i$ ,  $\Delta(u_i)$  in the tree  $T$  is that in  $T'_u$  plus  $\lambda^i \cdot \Delta(u)$ . Thus, we have

$$\begin{aligned} \sum_{v \in V(T) \setminus r} \Delta^2(v) &\leq \mathcal{V}(\rho n) + \mathcal{V}((1 - \rho)n) + \sum_{i=0}^r \left( (\Delta(u_i) + 2\lambda^i \Delta(u))^2 - \Delta^2(u_i) \right) \\ &= \mathcal{V}(\rho n) + \mathcal{V}((1 - \rho)n) + 4\Delta(u) \sum_{i=0}^r \lambda^i \Delta(u_i) + 4\Delta^2(u) \sum_{i=0}^r \lambda^{2i} \end{aligned}$$

From Claim 1, we can bound  $\Delta(u_i)$  by  $O(n^{1-\eta})$  for  $i = 0, \dots, r$ . So we get the following recurrence for  $\mathcal{V}(n)$

$$\mathcal{V}(n) \leq \mathcal{V}(\rho n) + \mathcal{V}((1 - \rho)n) + O(n^{2-2\eta})$$

which evaluates to  $\mathcal{V}(n) = O(n^{2-2\eta})$ .  $\square$

Lemma 2 and 3 prove the theorem.  $\square$

**Corollary 1.** If  $\mathbf{E}[Z] = \Omega(n^{1-\eta+\varepsilon})$  for any  $\varepsilon > 0$ , then  $\mathbf{P}[Z \in (1 \pm \rho)\mathbf{E}[Z]] \geq 1 - \frac{1}{\text{poly}(n)}$ .

*Proof.* Follows from a direct application of Chebyshev's inequality.  $\square$

## 4 Detecting Dependence

### 4.1 General Multistate Characters

Fix characters  $i$  and  $j$ . For brevity's sake, let  $N_e := M_{e,i} \otimes M_{e,j}$ . Recall by assumption, we have  $\|N_e\| \leq \lambda < 1/2$ . Let  $X$  be the compound character taking values in  $\mathcal{S} \times \mathcal{S}$  with  $X(u) := (X_i(u), X_j(u))$ . By our model of dependence, if  $i$  and  $j$  are dependent, then the compound character



evolves as a tree Markov random field process with transition matrices  $\hat{M}_e$  on each edge  $e$ , satisfying the property  $\hat{M}_e - N_e = \mathbf{1}\delta^T$  where  $\|\delta\|_1 = 2\delta > 0$  and  $\sum_{s \in \mathcal{S} \times \mathcal{S}} \delta(s) = 0$ . Since  $\|\delta\|_1 = 2\delta$ , we get there exists a subset  $S^* \subseteq \mathcal{S} \times \mathcal{S}$  such that  $\sum_{s \in S^*} \delta(s) = \delta$ . That is, dependent characters put an extra probability mass of  $\delta$  on the states in  $S^*$ , as compared to independent characters.

Fix a leaf  $\ell$  of the tree and let  $(e_1, \dots, e_t)$  be the path from root to  $\ell$ . Note that the matrices  $\hat{M} := \hat{M}_{e_1} \hat{M}_{e_2} \cdots \hat{M}_{e_t}$  and  $N := N_{e_1} N_{e_2} \cdots N_{e_t}$  denote the transition probability matrices for the compound character from root to leaf when the two characters are dependent and independent respectively. Our first claim shows that if each  $\hat{M}_e$  puts significantly more mass on the states in  $S^*$  than  $N_e$ , then so does  $\hat{M}$  over  $N$ .

**Lemma 4.** *Let  $E := \hat{M} - N$ . Then  $E$  has all rows equal, and the entries on the columns corresponding to  $S^*$  sum to  $\geq \lambda'\delta$ , where  $\lambda' = 1 - \frac{\lambda}{1-\lambda}$ .*

*Proof.* Let's start with a definition.

**Definition 1.** *A matrix is called an error matrix if all it's rows are equal and sum to 0.*

Note that the matrix  $D := \mathbf{1}\delta^T = \hat{M}_{e_i} - N_{e_i}$  for each  $i$ , is an error matrix. The following are a few easy to check properties of error matrices.

- Claim 2.** 1) *The sum of two error matrices is an error matrix.*  
2) *The product of two error matrices is the all zeros matrix.*  
3) *The product of a row stochastic matrix with an error matrix is the error matrix itself.*  
4) *The product of an error matrix with a row stochastic matrix is another error matrix.*

The above claim gives us,

$$\hat{M} = (N_{e_1} + D)(N_{e_2} + D) \cdots (N_{e_t} + D) = N + D + D(N_{e_1} + N_{e_1}N_{e_2} + \cdots + N_{e_1}N_{e_2} \cdots N_{e_{t-1}}),$$

This gives  $E = D + DN^*$ , where  $N^*$  is the sum of the matrices in the parentheses of the RHS above.  $N^*$  is also row stochastic, and furthermore,  $\|N^*\| \leq (\lambda + \cdots + \lambda^{t-1}) \leq \frac{\lambda}{1-\lambda}$ . Here we are implicitly using Lemma 1 and the fact that all norms are at most  $\lambda$ . Therefore,  $DN^*$  is an error matrix (and hence so is  $E$ ) where each row has  $\ell_1$  norm at most  $\frac{\lambda}{1-\lambda}2\delta$ . Thus, the coordinates of  $DN^*$  corresponding to  $S^*$  sum to at least  $-\frac{\lambda}{1-\lambda}\delta$ . This follows from the simple fact:

**Fact 1.** *Given any vector  $\delta \in \mathbb{R}^n$  with  $\sum_i \delta(i) = 0$  and  $\|\delta\|_1 \leq 2\varepsilon$ , then  $\max_{S \subseteq [n]} |\sum_{i \in S} \delta(i)| \leq \varepsilon$ .*

Therefore, the sum of the columns of each row of  $E$  corresponding to  $S$  is at least  $\delta(1 - \frac{\lambda}{1-\lambda})$ .  $\square$

Note that if  $\lambda < 1/2$ , we get that  $\lambda'$  is a constant bounded away from 0. Now we are ready to state our algorithm for detecting dependence in the multistate character case. The idea is pretty simple: if  $i$  and  $j$  are dependent, we expect a much more frequent occurrence of the states in  $S^*$  in  $\{(X_i(\ell), X_j(\ell)) : \ell \in L(T)\}$ , than what is expected if they were independent. One can exhaustively search for this set  $S^*$  since the number of states is a constant; the analysis follows almost immediately as a corollary of Theorem 1.

**Algorithm Multistate Dependence Detection.**

Input: States of the characters at  $n$  leaves,  $i, j$

Parameter: Precision parameter  $\varepsilon'$

1. Search for the special set  $S^* \subseteq \mathcal{S} \times \mathcal{S}$  exhaustively by going over all  $2^{s^2}$  (recall  $s$  is a constant).
2. Let  $\hat{Z}$  denote the fraction of leaves  $\ell$  such that  $(X_i(\ell), X_j(\ell)) \in S^*$ .
3. Let  $Z_a$  denote the fraction of leaves in state  $a$ , and let  $Z = \sum_{(a,b) \in S^*} Z_a Z_b$ .
4. If  $\hat{Z} \geq (1 - \varepsilon')(Z + \lambda'\delta)$ , declare  $i, j$  dependent. Else, declare  $i, j$  independent.

**Theorem 5** (Implies Theorem 2). *Algorithm Multistate Dependence Detection is correct.*

*Proof.* (Sketch) Suppose  $i$  and  $j$  are dependent. We show our algorithm declares them to be dependent whp. The proof when  $i$  and  $j$  are independent is similar. Note that if  $i$  and  $j$  were independent, then  $\mathbf{E}[Z] = \sum_{(a,b) \in S^*} \mathbf{E}[Z_a] \mathbf{E}[Z_b]$ , and thus by Lemma 4 we get  $\mathbf{E}[\hat{Z}] \geq \mathbf{E}[Z] + \lambda'\delta$ . By Corollary 1, we have that  $\hat{Z} \geq (1 - \varepsilon)(\mathbf{E}[Z] + \lambda'\delta)$  whp.

Furthermore, note that whenever  $\mathbf{E}[Z_a] = \Omega(1)$ , Corollary 1 gives  $\mathbf{E}[Z_a] \geq Z_a/(1 + \varepsilon)$  whp. This implies that even if  $\mathbf{E}[Z_a] \mathbf{E}[Z_b] = \Omega(1)$  for any  $(a, b) \in S^*$ , we get whp that  $\mathbf{E}[Z] \geq \frac{1}{1+\varepsilon} \sum_{(a,b) \in S^*} Z_a Z_b$ . If all of the  $\mathbf{E}[Z_a] \mathbf{E}[Z_b] = o(1)$ , and so  $\mathbf{E}[Z] = o(1)$  (finitely many characters), then Markov will imply that whp  $Z \ll \varepsilon \lambda' \delta$ . In either case, we get whp  $\hat{Z} \geq \frac{1-\varepsilon}{1+\varepsilon} Z + (1 - \varepsilon) \lambda' \delta \geq (1 - \varepsilon')(Z + \lambda' \delta)$  for the proper choice of  $\varepsilon'$ . To make the proof precise one needs to make  $o(1)$  precise; we avoid this in this abstract. □

## 4.2 Symmetric, two state CFN Model

For the case of two state, symmetric CFN model, we can analyze the counting algorithm similar to that of the previous subsection with a much weaker assumption. Note that the joint transition matrix in this case is a  $4 \times 4$  matrix. The reason we are able to obtain a better analysis in the 2-state symmetric model is because we can map any  $4 \times 4$  ‘history independent’ transition matrix onto a two state Markov chain. Two state Markov chains are much more well-behaved structures than those in higher states. We elucidate below.

By symmetry, we can assume that the special state preferred by the dependent characters is 00. Consider the three partitions of the state space:

$$\begin{aligned} \aleph &= \{(00), (11)\}; \beth = \{(01), (10)\}, \text{ or} \\ \aleph &= \{(00), (01)\}; \beth = \{(10), (11)\}, \text{ or} \\ \aleph &= \{(00), (10)\}; \beth = \{(01), (11)\} \end{aligned}$$

Note that for one of the three choices, we have  $|\sum_{s \in \aleph} \delta(s)| > 2\delta/3$ . By symmetry, suppose  $\aleph$  is such that the sum is positive. The reason for choosing the partitions is the following. We claim that any  $4 \times 4$  transition matrix of the form  $\hat{M}_e = M_{e,i} \otimes M_{e,j} + \mathbf{1} \delta^T$  can be mapped onto a transition matrix  $M'_e$  on two states  $\{\aleph, \beth\}$ , for any of the choices of  $\aleph, \beth$  above. It suffices to check that for each choice above that the probability of going from either state in  $\aleph$  to  $\beth$  is the same. For instance, suppose we are in the first partition. Then, we have

$$\hat{M}[00, 01] + \hat{M}[00, 10] = (1 - p)q + p(1 - q) + \delta(01) + \delta(10) = \hat{M}[11, 01] + \hat{M}[11, 10]$$

We are using crucially here the symmetry of the original  $2 \times 2$  transition matrix and the history independence. The other cases can be similarly checked. Also note that each probability is bounded away from 0 and 1 by constants depending only on  $\delta$  and  $\lambda$ . The following claim summarizes the discussion above.

**Claim 3.** *For any edge  $e$  and for any choice of  $\aleph, \beth$  above, there exists a well-defined transition matrix  $M'_e$  on the  $\{\aleph, \beth\}$  which is induced by the transition matrix  $\hat{M}_e$ . Furthermore, there is a choice of  $\aleph, \beth$  such that  $M'_e = \begin{pmatrix} 1 - \hat{p}_e + \hat{\delta} & \hat{p}_e - \hat{\delta} \\ \hat{p}_e + \hat{\delta} & 1 - \hat{p}_e - \hat{\delta} \end{pmatrix}$  for some  $\hat{p}_e \geq C_\lambda$ , and some  $\hat{\delta} \geq 2\delta/3$ .*

We are now ready to state the algorithm.

**Algorithm CFN Dependence Detection.**

Input: States of the characters at  $n$  leaves.

Parameter: Constant  $C'_{\lambda, \delta}$ , defined later.

For the three choices of  $\aleph$  and  $\beth$  defined above, evaluate the fraction of leaves whose  $i$ th and  $j$ th characters are in one of the states in  $\aleph$ . If for any one of them, this fraction is larger than  $\frac{1}{2} + 0.1C_{\lambda, \delta}$ , declare dependent. Else, declare independent.

**Theorem 6** (Implies Theorem 3). *Algorithm CFN Dependence Detection is correct.*

*Proof.* (Sketch) Let  $Z$  denote the number of leaves in state  $\aleph$ . If characters  $i$  and  $j$  are independent, then the transition matrix on any edge  $e$  is the tensor product of two symmetric matrices and hence symmetric. Therefore the stationary matrix is the all quarters vector. Since most leaves are at a distance  $\geq \log n$  from the root, and the chain is rapidly mixing ( $\lambda < 1$ ), we get  $\mathbf{E}[Z] \in (1/2 \pm o(1))n$ . By Corollary 1, whp  $Z \leq (1/2 + \varepsilon')n$  for any  $\varepsilon' > 0$ . Thus, whp independent characters are returned as independent.

Suppose  $i$  and  $j$  are dependent, and suppose  $\aleph, \beth$  is the partition which achieves the transition matrix as in Claim 3. Fix a leaf  $\ell$  and let  $(e_1, \dots, e_t)$  be the path from root to  $\ell$ ; let  $M' = M'_{e_1} \cdots M'_{e_t}$ . Note that if  $t$  is large enough,  $M'(\aleph, \aleph) \approx M'(\beth, \aleph)$ . It suffices to show that this quantity is bounded away from  $1/2$  by a quantity  $C'_{\lambda, \delta} > 0$ . This implies  $\mathbf{E}[Z] \geq (1/2 + C'_{\lambda, \delta})n$ , and the proof follows via Corollary 1.

Given a  $2 \times 2$  row stochastic matrix  $M = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$  with  $q \geq p$ , define the asymmetry  $\rho(M) = q/p$ .

**Fact 2.** *Given two row stochastic matrices  $M_1$  and  $M_2$ ,  $\rho(M_1 M_2)$  lies between  $\rho(M_1)$  and  $\rho(M_2)$ .*

*Proof.* If the asymmetry factor of  $M_i$  is  $q_i/p_i$ , then that of  $M_1 M_2$  is  $\frac{(q_1 + q_2 - q_1(p_2 + q_2))}{(p_1 + p_2 - p_1(p_2 + q_2))}$  which lies between  $q_1/p_1$  and  $q_2/p_2$ .  $\square$

From the above fact, we get that  $\rho(M') \geq \min_i \rho(M'_{e_i})$ . Since each  $M'_{e_i}$  has asymmetry at least  $\frac{\hat{p}_e + \hat{\delta}}{\hat{p}_e - \hat{\delta}}$ , we get that  $\rho(M')$  is bounded away from 1. This implies that  $M'(\beth, \aleph)$  is bounded away from  $1/2$  by a constant  $C'_{\lambda, \delta}$  only depending on  $\lambda$  and  $\delta$ .  $\square$

**Acknowledgements.** We would like to thank Elchanan Mossel and Nikhil Srivastava for useful discussions.

## References

- [1] J.A. CAVENDER. Taxonomy with Confidence. *Math. Biosci.*, **40**: 271–80 (1978).
- [2] J. T. CHANG. Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. *Mathematical Biosciences*, 137, 51–73, 1996.
- [3] C. DASKALAKIS, E. MOSSEL, AND S. ROCH. Optimal phylogenetic reconstruction. *Proc. 38th ACM STOC*, 159–166 (2006).
- [4] P.L. ERDÖS, M. STEEL, L. SZEKELY AND T. WARNOW. A few logs suffice to build (almost) all trees (I). *Random Structure and Algorithms*, 14, 153–184, 1997.
- [5] P.L. ERDÖS, M. STEEL, L. SZEKELY AND T. WARNOW. A few logs suffice to build (almost) all trees (II). *Theoretical Computer Science*, 221 (1–2), 77–118, 1999.
- [6] W. EVANS, C. KENYON, Y. PERES, AND L. SCHULMAN. Broadcasting on trees and the Ising model. *Annals of App. Prob.*, 10(2), 410–433, 2000.
- [7] J.S. FARRIS. A probability model for inferring evolutionary trees. *Syst. Zool.* **22**:250–56 (1973).
- [8] M. FARACH AND S. KANNAN. Efficient algorithms for inverting evolution. *Proc. 28th ACM STOC*, 1996.
- [9] J. FELSENSTEIN. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–76 (1981).
- [10] J. FELSENSTEIN. *Inferring Phylogenies*. Sinauer, New York, 2004.
- [11] Y. HIGUCHI. Remarks on the limiting Gibbs states on a  $(d + 1)$ -tree. *Publ. RIMS, Kyoto Univ.* 13, 335–348, 1977.
- [12] J. HUELSENBECK AND K. CRANDALL. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.* **28**:437–66 (1997).
- [13] D. A. LEVIN, Y. PERES, AND E. L. WILMER Markov Chains and Mixing Times. *American Mathematical Society*, ISBN-10: 0-8218-4739-2, 2008.
- [14] W. MADDISON. A Method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution*, **44**(3), 539–557 (1990).
- [15] E. MOSSEL. Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356**:6 2379–2404 (electronic) 2004.
- [16] E. MOSSEL AND S. ROCH. Learning Nonsingular Phylogenies and Hidden Markov Models. *Proc. of 37th ACM STOC*, 2005.
- [17] J. NEYMAN. Molecular studies of evolution: a source of novel statistical problems. In *Statistical decision theory and related topics*, S.S. Gupta and J. Yackel (eds.) 1–27 (1971).

- [18] C. SEMPLE AND M. STEEL. *Phylogenetics*,. Oxford Lecture Series in Mathematics and its Applications. **24**.
- [19] F. SPITZER. Markov Random Fields on an Infinite Tree. *Annals of Prob.*, 3(3), 387–398, 1975.
- [20] N. SRIVASTAVA. Personal communication.
- [21] D.L. SWOFFORD, G.J. OLSEN, P.J. WADDELL, AND D.M. HILLIS. Phylogenetic Inference. In: *Molecular Systematics*, 2nd edn. (ed. D.M. Hillis, C. Moritz, and B.K. Marble). Sinauer, Sunderland, USA 407–514 (1996).